Comparison of Predictive Modelling Methods in Classification Problems

Sujata Suvarnapathaki

Ramnarain Ruia Autonomous College, sujatasuvarnapathki@riuiacollege.edu

Abstract: Predictive modelling in data science is a growing area. Historical data are used to develop predictive models. Predictive models can help to manage the resources efficiently. Classical statistical modelling technique ‘Binary Logistic Regression’ is primarily used to develop predictive models when the response variable is binary. Binary logistic regression models, the probability of success conditioning on the set of factors/independent variables and is one of the most preferred methods used in Classification Problem. Other methods such as the Naïve Bayes classifier and Random Forest are being used as an alternative to Binary Logistic Regression. This paper compares binary logistic regression, Naïve Bayes method and Random Forest method. All these methods were performed on the open data sets. These methods were compared using the AUC, area under the ROC curve (Receiver Operating Characteristic curve) obtained on the validation dataset called, the test data. ROC curve plots sensitivity vs. 1-specificity for various probability cut-offs. Independent variables were selected using the domain expertise and the same set of variables was used in all the methods.

Although Binary Logistic Regression remains the most preferred modelling technique, other methods may outperform in specific cases. It is advisable to check the performance of other modelling techniques before the model is implemented.

Index Terms: Classification Problem, Predictive Modelling, Binary Logistic Regression, Naïve Bayes Method, Random Forest Method.

I INTRODUCTION:

Predictive modelling is a process to create a statistical model to best predict the outcome or probability of an outcome.

Predictive analytics is used in Market research, Human Resources industry, Banking industry, Healthcare industry, financial services, Insurance, Telecommunications and other fields. The statistical model in these industries can be a classification or a regression model.

In the case of a classification model, the class probabilities are estimated whereas the mean values are estimated in the regression problem.

Classical statistical modelling technique, ‘Binary Logistic Regression’, is primarily used to develop predictive models when the response variable is binary.

Binary logistic regression models, the probability of success conditioning on a set of factors/independent variables.

Other methods such as Naïve Bayes and Random Forest are used as an alternative to Binary logistic regression.

This paper compares the Binary Logistic Regression, the Naïve Bayes method and the Random Forest method using the open datasets from the banking industry, market research, human resources industry.

The models were developed on the training datasets and prediction accuracy was compared using validation datasets, called test data.

II METHODOLOGY:

Binary Logistic Regression, Naïve Bayes Method and Random Forest method were applied on the open data sets and the results are compared.
All these three methods were used to classify a ‘case’ into one of the two categories: ‘success’ or ‘failure’. The response variable ‘Y’ was binary whereas independent variables were a mix of categorical and continuous in all data sets used for comparison.

Figure 1:

![Diagram showing Binary Dependent Variable or response variable and Independent variables-mix of categorical and continuous variables.]

### a) Binary Logistic Regression:

Binary Logistic Regression is the most widely used statistical modelling technique in business and research.

The mathematical model in Logistic Regression is given below:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k$$

where \(p\) is the probability of an event conditioning upon values of independent variables. \(p = P(Y/X_1, X_2, \ldots, X_K)\), where \(Y\) is the binary dependent variable and \(X_1, X_2, \ldots, X_K\) are independent variables. The Method of Maximum Likelihood Estimation is used to estimate the parameters of the model. Two commonly used iterative maximum likelihood procedures are the Fisher scoring method and the Newton-Raphson method. The parametric tests of significance are performed to test the overall significance of the model as well as the significance of the individual explanatory variables.

Receiver Operating Characteristic curve (ROC Curve), is a tool for assessing the performance of the classification model. For the predictive model, ROC Curve uses different probability values for which a summary of the trade-off between the true positive rate and false-positive rate is obtained. This summary is represented as the Area Under the Curve (AUC). The AUC measures the ability of a classifier to distinguish between classes. The high value of the AUC indicates better performance of the model at distinguishing between the classes. The large value of AUC (>0.80) indicates an excellent performance of the model. However, the model fails if the value of AUC ranges between 0.50 to 0.60 or if it is less than 0.50.

### b) Naïve Bayes Method:

The Naïve Bayes method is a supervised Machine Learning algorithm. It is used in the Classification Problem. Therefore, it can be considered as an alternative to Binary Logistic regression as well as Multinomial Logistic Regression. Naïve Bayes classification method is based on the Bayes theorem of conditional probability. The conditional probability, \(P(Y/X_1, X_2, \ldots, X_K)\) is obtained for each category of the response variable, \(Y\) and an observation is classified using the maximum conditional probability.

The Naïve Bayes Classifier assumes strong conditional independence among the predictors. The Naïve Bayes’ method is suitable when the number of input variables is high. Though it is a very simple method, in some specific situations it may outperform more sophisticated classification methods. Naïve Bayes method can estimate the parameters required for the classification using small training data. The evaluation of the classifier is quick and easy.

The method can be a good alternative to logistic regression. The performance of the Naïve Bayes method can be assessed using the ROC Curve and the area under the curve (AUC).

### c) Random Forest Method:

Random forest is a classification method that comprises of many decision trees. ‘Ensemble Learning’ methods generate multiple classifiers. The results of these classifiers are aggregated. Boosting and Bagging are the two well-known methods of classification trees. The successive trees give extra weight to the points incorrectly predicted by the earlier predictors in Boosting. A weighted vote is taken in the end, for prediction. Breiman introduced the term “Bagging” which stands for “bootstrap aggregating”. It is an ensemble method: a method of combining results from multiple resamples. The Ensemble method can also be applied by using different classifiers for a given sample. Each tree is independently constructed using a bootstrap sample of the dataset. A simple majority vote is taken for prediction in the end. Later in 2001, Breiman proposed random forests, which add an extra layer of randomness to bagging. The method constructs each tree using a different bootstrap sample of the data. The random forests change how the classification or regression trees are constructed. Each node is split using the best split among all variables in standard trees whereas in a random forest, each node is split using the best among a subset of predictors randomly chosen at the node. The final output is the class that is the mode of the class's output by the individual trees.
Figure 2:

The Random Forest algorithm works as follows a) First, grow a forest of many trees. (R default is 500), b) grow each tree on an independent bootstrap sample (Sample N cases at random with replacement) from the training data c) At each node: select m variables at random out of all M possible variables (independently for each node), find the best split on the selected m variables d) grow the trees to maximum depth (classification) e) vote/average the trees to get predictions for new data. The performance of the Random Forest method can be assessed using the ROC Curve and the area under the curve (AUC).

Four open data sets were used to compare the three methods above. Independent variables were selected using domain expertise and the same set of variables was used in all the methods. The problem of Multicollinearity was handled at the time of variable selection.

Models were developed using a random sample from an open Dataset, called train data and scoring was performed on the entire data. For each data, all the three methods were applied on the same Train Data and Test Data. For each method, ROC Curve was plotted and the area under the ROC Curve was obtained. ROC curve plots sensitivity vs. 1-specificity for various probabilities cut-offs. It summarizes the trade-off between true-positive rate and false-positive rate.

In the end, the overall model performance was evaluated using the ‘ROC curve and the area under the ROC curve (AUC) for each method.

An open-source statistical programming software ‘R’ was used to develop models for the three methods described above for the datasets under consideration.

The first dataset was from the Banking Industry with the objective of predicting the customer churn. The dataset had 10,000 observations. There were 10 predictors in the dataset. This data set contained the details of a bank’s customers who were withdrawing(closing) their account from the bank due to some loss and other issues. The response variable was a binary variable indicating whether the customer left the bank (closed his account) or continued to be a customer.

The second dataset was from the healthcare industry. This dataset is available on the website www.kaggle.com. The dataset provides the patients’ information regarding their cardiovascular conditions. It includes 4,238 records and 15 attributes.

The objective was to classify if a patient had a 10-year risk of future coronary heart disease.

The third dataset was from the Human Resources industry. For any organization, the problem of employee retention and the recruitment of new employees is crucial. The knowledge about the potential employee churn would be very helpful in this regard. This falls under popularly called “HR (Human Resources) Analytics”. This is a standard supervised classification problem. The dataset was related to Employee turnover (also known as “employee churn/attrition”). In this case, the target variable Y, was to predict if an employee would leave the company. This dataset included 1470 records and 29 independent variables.

The fourth dataset was regarding ‘water potability’. Drinking safe water is essential to good health. It is a component of effective policy for health protection. The safety of drinking water could be a national, regional or local level important issue. This fourth dataset is an open dataset available on the website www.kaggle.com. This data consisted of water quality metrics for 3276 different water bodies and there were 9 predictors. The objective was to classify the water sample to be ‘potable’ or not.

III RESULTS AND DISCUSSION:

In the first case, the objective was to classify a customer as ‘likely to churn’ or not. There were 10,000 cases and 10 explanatory variables of which 5 were categorical and 5 were continuous variables. The data were randomly divided into Train data, with 7000 observations and the Test data with 3000 observations. The values of the AUC for ROC Curves for Train data and Test data were obtained for Binary Logistic Regression, Naïve Bayes method and Random Forest method. The values of the AUC for the ROC curves for the ‘Train data’ were observed to be 0.7679, 0.8002 and 1 respectively. The AUC for ROC Curve for the ‘Test data’ was 0.7607 for Binary Logistic Regression. The AUC for ROC Curve for the same Test data was 0.7935 for the Naïve Bayes method and it was 0.8412 for the Random Forest method.

In this case, the Naïve Bayes method outperformed the Binary Logistic Regression whereas the Random Forest method outperformed both, the Binary Logistic Regression and the Naïve Bayes method.

The second case was about the heart attack prediction. The classification goal was to predict whether the patient had a 10-year risk of future coronary heart disease (CHD). The dataset
provided the patients’ information. It included 4,238 records of which 3256 were non-missing records. There were 15 explanatory variables, of which 6 variables were categorical and 9 variables were continuous. The data were randomly divided into Train data, with 2560 observations and the Test data with 1096 observations. The values of the AUC for the ROC Curves for Train data and Test data were obtained for the Binary Logistic Regression, the Naïve Bayes method and the Random Forest method. The values of the AUC for the ROC curves for the ‘Train data’ were observed to be 0.7312, 0.7244 and 1 respectively. The AUC for ROC Curve for the ‘Test data’ was 0.7460, using the Binary Logistic Regression. The AUC for ROC Curve for the same Test data was 0.7263 for the Naïve Bayes method and it was 0.7164 for the Random Forest method.

In this case, the Binary Logistic Regression outperformed the Naïve Bayes method and the Random Forest method whereas the Naïve Bayes method outperformed the Random Forest method.

In the third case, an open dataset on employee attrition in Human Resource (HR) Analytics domain was used. The response variable “Attrition” was a binary variable and there were 29 explanatory variables of which one variable had a severe multicollinearity problem. This variable was dropped and 28 variables were included for the analysis. The data were randomly divided into Train data with 1029 observations and the Test data consisting of 441 observations. The values of the AUC for ROC Curves for Train data and Test data were obtained for the Binary Logistic Regression, the Naïve Bayes method and the Random Forest method. The values of the AUC for the ROC curves for the ‘Train data’ were 0.8595, 0.8323 and 1 respectively. The AUC for ROC Curve for the ‘Test data’ was 0.8727, for the Binary Logistic Regression and was 0.8448 for the Naïve Bayes method and the value of the AUC for the ROC for the Test data was 0.8610 for the Random Forest method.

In this case, it is observed that the Binary Logistic Regression outperformed the other two methods.

In the fourth case, the dataset for water potability was used. It consisted of 3276 observations but only 2011 non-missing cases. There were 9 explanatory variables. The target variable was a binary variable indicating whether the water was ‘potable’ or ‘not’. No multicollinearity was observed. The data were randomly divided into Train data, with 1408 observations and the Test data consisting of 603 observations. The values of the AUC for ROC Curves for Train data and Test data were obtained for the three methods, viz. Binary Logistic Regression, Naïve Bayes method and Random Forest method. The values of the AUC for the ROC curves for the ‘Train data’ were observed to be 0.5316, 0.6354 and 1 respectively. The AUC for ROC Curve for the ‘Test data’ was 0.5129, using the Binary Logistic Regression. The AUC for ROC Curve for the same Test data was 0.6179 under the Naïve Bayes method and the value of the AUC for the ROC for the Test data was 0.7116 for the Random Forest method.

In this case, the Binary Logistic Regression completely failed as it had the lowest value for AUC. Naïve Bayes method was observed to be better than the Binary Logistic Regression but the Random Forest Method outperformed both the Binary Logistic regression and the Naïve Bayes method.

Here is the summary table indicating the AUC for three methods.

<table>
<thead>
<tr>
<th>No.</th>
<th>Dataset with the objective</th>
<th>Number of Observations</th>
<th>Number of Observations Test Data</th>
<th>Binary Logistic Regression AUC For Test Data</th>
<th>Naïve Bayes Method AUC For Test Data</th>
<th>Random Forest Method AUC For Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Customer Churn prediction</td>
<td>7000</td>
<td>3000</td>
<td>0.7607</td>
<td>0.7935</td>
<td>0.8412</td>
</tr>
<tr>
<td>2</td>
<td>Heart Attack prediction</td>
<td>2560</td>
<td>1096</td>
<td>0.7460</td>
<td>0.7263</td>
<td>0.7164</td>
</tr>
<tr>
<td>3</td>
<td>Employee Attrition prediction</td>
<td>1029</td>
<td>441</td>
<td>0.8727</td>
<td>0.8448</td>
<td>0.8610</td>
</tr>
<tr>
<td>4</td>
<td>Water Potability prediction</td>
<td>1408</td>
<td>603</td>
<td>0.5129</td>
<td>0.6179</td>
<td>0.7116</td>
</tr>
</tbody>
</table>

### IV CONCLUSION:

Although the Binary Logistic Regression remains the most preferred modelling technique, other methods may outperform in specific cases. It was observed that the Naïve Bayes method and the Random Forest outperformed the Binary Logistic Regression in one of the four cases. It is suggested that even if the most preferred technique of classification fails, the classification objectives can still be met by applying other techniques. It is recommended to check the performance of other modelling techniques before implementing the Binary Logistic regression model. Further research can be done by simulating hundreds of data sets and comparing the above methods using the area under the ROC Curve.

### REFERENCES:


DATA SOURCE:
https://www.kaggle.com

***